JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 9, pp. 1397-1407, Sep. 2021

Check for updates

딥러닝 기반 멀티모달 정보 분석을 활용한 라이브커머스 영상 요약 기법

황 중 수[1] · 조 영 선[1] · 박 상 훈[1] · 이 현 기[1] · 손 종 수[2*]
[1]CJ올리브네트웍스 AI Core 연구소 연구원
[2*]CJ올리브네트웍스 AI Core 연구소 연구소장

Live-commerce Video Summarization using Analysis of Multi-modal Information based on Deep Learning

Jung-Su Hwang[1] · Young-Sun Cho[1] · Sang-Hoon Park[1] · Hyun-Ki Lee[1] · Jong-Soo Sohn[2*]

[1]Researcher, AI Core Research Center, CJ Olivenetworks, 366, Hangang-daero, Yongsan-gu, Seoul, Korea
[2*]Director, AI Core Research Center, CJ Olivenetworks, 366, Hangang-daero, Yongsan-gu, Seoul, Korea

[요 약]

라이브커머스 서비스는 코로나바이러스감염증-19(COVID-19)로 인한 언택트 서비스의 부상과 함께 급성장하고 있는 분야 중 하나이다. 라이브커머스에는 실시간 방송 서비스와 지난 방송을 보기 위한 주문형 비디오(VOD; Video On Demand) 서비스가 있다. 본 논문에서는 VOD 서비스에서 지난 방송의 제품 정보, 이벤트와 같은 주요 내용이 포함된 요약본을 제공하기 위해 딥러닝 기반으로 영상에서 임팩트 있는 구간만 자동으로 간추리는 영상 요약 기법을 제안한다. 라이브커머스 영상이 담고 있는 오디오, 이미지, 텍스트로 구성된 멀티모달 정보를 딥러닝 기반으로 분석하여 구간별 임팩트 스코어를 측정하고, 각 멀티모달 스코어를 합산하여 임팩트가 강한 구간만 추출하는 방식이다. 본 기법으로 라이브커머스의 특징을 담고 있는 임팩트 강한 요약 영상을 자동 생성함으로써 요약본 제공의 적시성을 확보하고 영상 편집에 필요한 시간과 비용을 절감한다.

[Abstract]

Recently, Live-commerce is one of the services that are growing rapidly with untact service by COVID-19. Live-commerce includes a real-time service and Video On Demand(VOD) service about past broadcasts. In this paper, we propose video summarization technique using Deep Learning for providing summarized video including main contents such as product information or events and so on. This is for offering high impact of past broadcast. After multi-modal information composed of Audio, image, text is analyzed by Deep Learning, we extract impact score per section about each multi-modal information. Section with high impact score is selected by summing each multi-modal score. It reduces the editing time and costs required for summarizing video so that retains proper timing by creating the effective summary video containing characteristics of Live-commerce automatically.

색인어 : 딥러닝, 라이브커머스, 멀티모달, 영상 처리, 영상 요약

Keyword : Deep learning, Live-commerce, Multimodal, Video Processing, Video Summarization

http://dx.doi.org/10.9728/dcs.2021.22.9.1397

Received 24 August 2021; **Revised** 07 September 2021
Accepted 07 September 2021

***Corresponding Author; Jong-Soo Sohn**
Tel: ※ 개인정보 표시제한
E-mail: jongsoo.sohn@cj.net

Ⅰ. 서 론

라이브커머스는 COVID-19로 비대면 쇼핑이 활발해지면서 급격하게 성장하고 있다. 라이브커머스는 전통적인 TV 홈쇼핑과는 다르게 주로 모바일 플랫폼에서 서비스되므로 쇼호스트나 판매자, 인플루언서와 시청자가 실시간으로 교감하며 재미를 유발함과 동시에 즉시성 있게 상품을 판매하는 장점이 있다. 이처럼 실시간으로 방송되는 서비스와 더불어 방송 종료 이후에도 판매의 연속성을 갖고 고객이 쇼핑하듯 지난 판매 상품을 골라볼 수 있는 VOD 서비스 또한 필요하다.

라이브커머스 방송은 대개 한 시간 이상 진행된다. 그러나 VOD 서비스를 원본과 같은 길이로 제공하면 라이브 방송만큼의 임팩트를 주기 어렵다. 따라서 영상의 지루함을 없애고 방송 원본에 대한 시청자의 관심을 유도하기 위해 임팩트 있는 구간만 짧게 요약할 필요가 있다. 다만, 영상 요약은 편집자의 노동력과 주관이 개입되며, 종료된 방송을 편집하여 요약방송으로 재가공하기에는 일정 시간 이상을 필요로 하는 한계점이 있다.

본 논문에서는 딥러닝 기반 멀티모달 정보 분석을 통해 영상에서 임팩트가 강한 구간을 찾아내 영상을 요약하는 방법을 제안한다. 영상에서 추출할 수 있는 멀티모달 정보는 오디오, 이미지, 텍스트가 있다. 첫 번째로 오디오에서 일반적인 톤의 발화 이외의 특이음(효과음, 웃음소리, 높은 톤의 음성 등)이 있는 구간을 찾고, 두 번째로 얼굴과 손이 클로즈업되는 구간을 찾는다. 세 번째로 대사에서 긍정적인 문장으로 발화하는 구간을 찾는다. 마지막으로 구간 추출 알고리즘을 통해 세 가지 멀티모달 정보에서 추출된 구간을 조합하여 영상에서 임팩트 강한 구간만 요약한다. 본 논문에서 제안하는 방법으로 방송이 종료됨과 동시에 자동으로 요약본을 단시간 내에 생성할 수 있고, 시청자들은 많은 영상을 임팩트 강한 구간만 빠르게 둘러볼 수 있다.

Ⅱ. 관련 연구

최근 몇 년간 영상 요약을 통해 원본 영상의 주요한 부분을 담아내고자 하는 연구가 다수 진행되었다. Zhou et al.[1]은 강화학습 기반으로 원본을 간략하게 표현할 수 있는 비디오 요약 기법을 제시하였다. 해당 기법은 요약된 영상이 원본 영상에 대해 다양성과 대표성을 나타내도록 reward를 구성하였다. Rochan and Wang [2]은 페어링 되지 않은 원본 영상과 요약 영상 데이터를 사용하여 원본 비디오와 요약 비디오 간의 매핑 관계를 GAN 기반으로 학습하는 비디오 요약 모델을 제안하였다.

하지만 위 연구들은 영상이 담고 있는 오디오, 텍스트, 이미지와 같은 멀티모달 정보 전체를 활용하기보다는 이미지 정보만 활용한 영상 요약 기법이라는 한계가 있다. 라이브커머스 영상은 주로 한정된 장소에서 소수의 인원으로 진행되고 정해진 상품만 주로 나오는 등 영상 자체에서 시각적인 변화가 크게 일어나지 않는 특징이 있다. 따라서 이미지 정보만 영상 요약에 활용할 경우 라이브커머스 특유의 특징을 살려내기 어렵다. 본 연구에서는 딥러닝 기술을 사용하여 영상의 이미지 정보뿐만 아니라 오디오 정보, 텍스트 정보를 함께 분석하고 도출된 분석 결과를 가지고 전체 영상에서 임팩트 있는 구간을 찾아내는 방식으로 라이브커머스 영상의 주요 특징들을 잘 담아낼 수 있는 영상 요약 기법을 제안한다.

Ⅲ. 멀티모달 분석 기반 영상 요약

라이브커머스 방송은 구매 증가를 위해 상호작용, 현장감, 상품의 상세정보 전달이 중요한 요소이다.[3] 본 연구에서는 오디오 분석을 통해 시청자와의 상호작용과 현장감이 높은 구간을 검출하고, 이미지와 텍스트 분석을 통해 상품의 상세정보 확인이 가능한 구간을 찾아 영상을 요약한다. 영상의 멀티모달 정보 분석으로 영상이 구간별로 가지고 있는 임팩트를 점수로 환산하여 점수가 높은 구간을 추출하는 방식이다.

멀티모달 정보에서 추출되는 스코어는 다음과 같다. 첫 번째로, 오디오 분석을 통해 높은 피치의 음성, 효과음과 같은 특이음들을 찾아내 구간별 Sound score를 측정한다. 두 번째로, 이미지 분석을 통해 영상 구간별로 얼굴과 손의 클로즈업 정도를 기반으로 Close-up score를 계산한다. 마지막으로 음성인식(STT; speech to text)을 통해 음성을 텍스트로 변환하고 텍스트 감정분석을 통해 대사의 긍정도를 측정하여 Positivity score를 측정한다. 도출된 세 개의 스코어를 기반으로 최종적으로 영상의 주요 구간을 선택하여 요약한다.

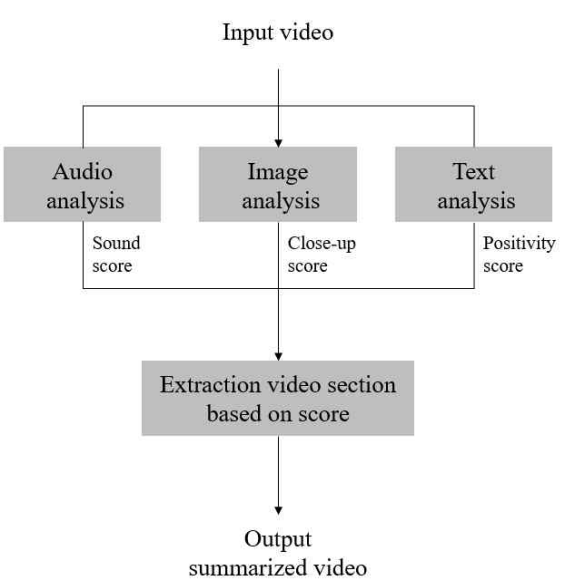

그림 1. 멀티모달 분석 기반 영상 요약 흐름도
Fig. 1. Video summarization flow by multi-modal information analysis

3-1 특이 음향 추출을 위한 음원 분리 모델 생성

라이브커머스 방송은 특유의 활기찬 분위기를 담고 있다. 따라서 일반적인 톤의 음성뿐만 아니라 이벤트 효과음, 박수 소리, 높은 톤의 음성과 같이 다양한 소리를 담고 있다. 이처럼 다양하게 인식되는 특이 음향의 데시벨 크기를 측정하여 Sound score로 사용한다.

본 연구에서는 Wave-U-Net[4]을 활용하여 입력된 오디오에서 특이음만 분리한다. Wave-U-Net은 audio signal separation 대회인 'SISEC 2018'에서 좋은 성능을 보인 end-to-end audio source separation 모델이다. Wave-U-Net은 U-Net 구조를 시간 도메인에 적용한 모델로, upsampling에서 발생하는 오디오 artifacts를 방지하기 위해 linear interpolation 사용을 제안한 모델이다. Wave-U-Net을 통해 입력된 오디오에서 원하는 오디오만 따로 분할 할 수 있다. 본 연구에서는 일반적인 톤의 음성과 효과음, 고음, 웃음소리와 같은 특이음의 추출할 목적으로 음원 분리를 진행한다. 원본 오디오를 입력하면 Wave-U-Net에 의해 일반 음성에서 특이음만 분리할 수 있고, 분리된 특이음을 대상으로 스코어를 계산한다. Sound score 산출을 위해 특이음을 일정한 간격으로 구분한다. 그리고 구간별로 오디오의 데시벨을 측정한 후 크기에 따라 Sound score를 계산한다.

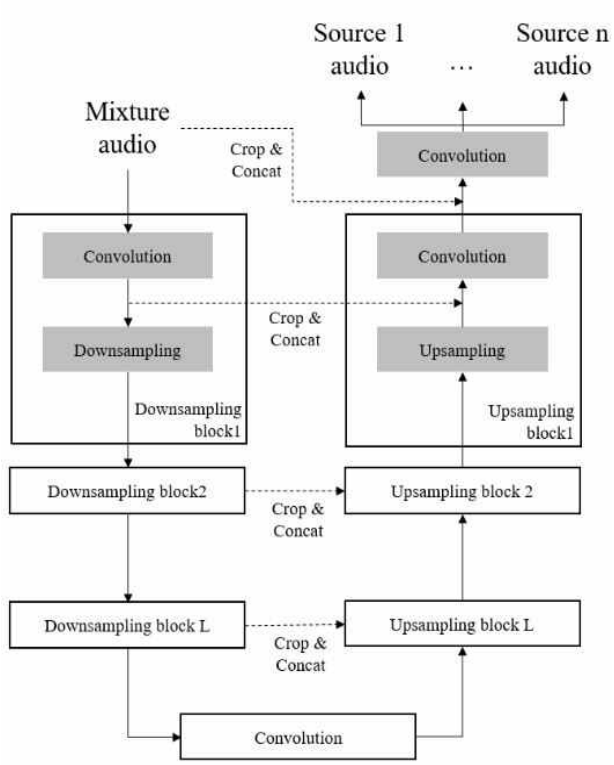

그림 2. Wave-U-Net 구조
Fig. 2. Wave-U-Net architecture

3-2 클로즈업 측정을 위한 프레임 내 얼굴, 손 검출 모델 생성

라이브커머스는 정면 카메라, 측면 카메라 등 다양한 구도에서 촬영이 진행된다. 특히 제품을 상세하게 보여주거나 제품을 시연할 때 주로 카메라가 클로즈업되고 이 부분은 방송에서 주요한 구간이 될 수 있다. object detection 기술을 활용하여 방송되고 있는 상품 자체가 클로즈업되는 구간을 주요 구간으로 찾을 수 있으나, 라이브커머스의 모든 판매 상품에 대한 데이터를 수집하기에는 한계가 있고, 포장을 뜯어 보여주는 경우와 같이 원형이 아닌 형태로 제품을 보여주는 경우가 있기 때문에 제품을 직접 인식하는 방법은 적용에 어려움이 있다. 따라서 본 논문에서는 상품을 소개할 때 주로 얼굴과 손이 이용되기 때문에 우회적인 방법으로 얼굴과 손이 클로즈업되는 구간을 잡아내 주요 구간으로 사용한다.

object detection을 위해 YOLO v3[5]을 사용한다. YOLO v3는 입력된 이미지 내에서 물체를 인식하고 해당 물체의 bounding box를 찾는 1-stage 기반 object detection 모델로, 객체에 대한 classification과 localization 문제를 동시에 해결하기 때문에 두 가지 문제를 순차적으로 해결하는 2-stage detector에 비해 속도가 빠르다. 라이브커머스 방송은 주로 한 시간 이상 진행되기 때문에 일부 프레임만 지정하여 분석하더라도 많은 양의 이미지 분석이 필요하다. 따라서 속도가 빠른 object detection 모델이 필요하고 이를 통해 방송 종료 후 빠르게 요약 영상을 생산해낼 수 있다.

$$\lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}[(x_i-\hat{x_i})^2+(y_i-\hat{y_i})^2] \quad (1)$$
$$+\lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w_i}})^2+(\sqrt{h_i}-\sqrt{\hat{h_i}})^2]$$

수식 1은 YOLO v3의 loss function 중 object의 bounding box를 예측하기 위한 localization loss 항으로 정답 좌표값인 (x,y)과 예측 좌표값인 $(\hat{x},\hat{y})$간의 차이와 정답 bounding box인 (w,h)과 예측 bounding box인 $(\hat{w},\hat{h})$간의 Loss를 계산함으로써 object의 위치와 영역을 검출해낼 수 있게 된다. 본 연구에서의 프레임에 얼굴과 손의 존재 여부를 확인하고, 검출된 bounding box의 크기를 기준으로 클로즈업 정도를 계산한다.

3-3 문장의 감정분석을 통한 긍정도 측정 모델 생성

출연자는 라이브 커머스 중 다양한 대화를 하고 그 중 상품에 대해 소개하거나 구체적으로 설명할 때 특히 긍정적인 어구들을 사용하게 된다. 예를 들어 쇼호스트가 특정 화장품에 대해 설명할 경우, '보습이 매우 뛰어납니다.', '발색이 정말 예뻐요.'와 같은 문장은 '색은 브라운 계열입니다.', '문지르듯이 바르면 돼요.'와 같은 문장보다 더욱 긍정적으로 표현된 문장이다.

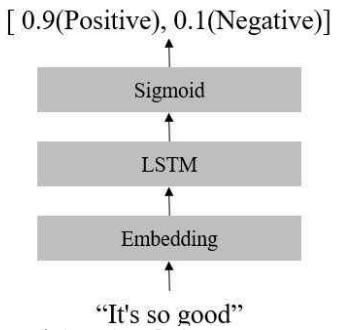

그림 3. 입력 문장의 긍정도 측정
Fig. 3. Calculating positivity of input sentence

$$sigmoid(x) = \frac{1}{1+e^{-x}} \quad (2)$$

본 연구에서는 binary classification을 위한 LSTM(Long Short-Term Memory) 기반의 모델을 통해 입력된 문장이 내포하고 있는 긍정도를 측정하는 방법을 제안한다.

그림 3과 같이 문장 단위로 입력되는 텍스트 데이터는 Embedding layer를 거치면서 벡터화되어 LSTM Layer에 순차적으로 입력된다. 마지막으로 Sigmoid 함수를 거치면서 입력된 문장은 긍정 또는 부정으로 분류된다. 수식 2와 같이 Sigmoid 함수의 결과값은 0에서 1 사이의 값을 가지게 된다. 따라서 binary classification 과정을 통해 입력된 문장이 Positive 클래스일 확률을 측정하고, 이를 문장의 Positivity score로 사용한다.

3-4 주요 구간 선정

3-1, 3-2, 3-3의 과정으로 추출된 세 가지 멀티모달 스코어를 모두 0~1 사이의 값으로 normalization하고, 합산하여 대상 구간 선정을 위한 최종 스코어를 계산한다.

주요 구간 선정 시 편집자의 의도, 요약본의 목적에 따라 구간이 다르게 선택되도록 각 score의 가중치를 다르게 하여 합산할 수 있다. 예를 들어 라이브커머스 방송 특유의 활기찬 분위기 또는 쇼호스트의 높은 텐션이 다수 포함되어있는 요약본을 생성할 때는 Audio score의 가중치를 높여 음향적인 특징이 두드러지는 영상을 뽑을 수 있다. 반면, 제품에 대한 소개나 시연이 많이 포함된 영상을 만들고자 할 경우, Close-up score의 가중치를 높게 줄 수 있다. 이처럼 편집자의 요약 의도에 따라 스코어별 가중치를 조절하여 특정 멀티모달 정보에 중점을 둔 요약 영상을 생성할 수 있다.

표 1과 같이 가중치 조절이 없는 경우, a 구간이 가장 높은 스코어를 가지고 있으나, 표 2처럼 Close-up score에 1.5배의 가중치를 주면 c 구간이 가장 높은 점수로 변경된다. 따라서 사용자의 목적에 따라 스코어의 가중치를 조절하여 주요 구간을 다르게 하여 영상을 요약할 수 있다.

표 1. 구간별 임팩트 점수 산정 예시
Table 1. Impact scoring example per video section

Section	Start Time(s)	Audio Score	Image Score	Text Score	Total Score
a	10	0.6	0	0.5	1.1
	11	0.6	0	0.5	1.1
b	12	0.1	0	0.5	0.6
	13	0	0.2	0.5	0.7
c	14	0	0.5	0.5	1
	15	0	0.5	0.5	1

표 2. 표1의 점수 대상, 가중치 적용 후 점수 재산정 예시(오디오 가중치 : 0.5, 이미지 가중치 : 1.5, 텍스트 가중치 : 1.0)
Table 2. Image score rescoring about Table 1. example (audio weight : 0.5, image weight : 1.5, text weight : 1.0)

Section	Start Time(s)	Audio Score	Image Score	Text Score	Total Score
a	10	0.3	0	0.5	0.8
	11	0.3	0	0.5	0.8
b	12	0.05	0	0.5	0.55
	13	0	0.3	0.5	0.8
c	14	0	0.75	0.5	1.25
	15	0	0.75	0.5	1.25

Ⅳ. 실 험

영상에서 Sound score, Close-up score, Positivity score를 추출하기 위해 다음 세 가지 분석 모델들을 생성하였다.

4-1 Sound score 추출 모델

음원 분리 학습을 위해 일반적인 톤의 음성과 그 외의 음향으로 구성된 데이터셋이 필요하다.

원본 영상의 오디오에서 효과음, 고음 발성과 같은 특이음과 일반적인 음성을 분리하기 위해 MUSAN 데이터셋을 활용하였다. MUSAN[6]은 music, speech, noise 음원으로 구성된 데이터셋이다. 본 실험에서는 특이음 분리를 위해 speech 음원과 noise 음원을 사용하여 모델을 학습하였다. speech 음원은 11개 언어로 구성된 약 60시간 길이의 음성 데이터이고, noise 음원은 발신음, 팩스음, 기계음, 천둥, 동물소리 등으로 구성된 약 6시간 길이의 데이터이다.

우선 speech 음원들과 noise 음원들을 랜덤하게 혼합하여 약 9시간 길이의 오디오를 생성하였다. 다음으로 Wave-U-Net이 혼합된 오디오 속에서 Noise 음만 추출하도록 학습하였고 그 결과, Wave-U-Net은 그림 4와 같이 입력된 오디오에서 일반 음성과 noise 음을 분리하게 된다. 결과적으로 라이브커머스 방송의 오디오에서 특이음들만 분리할 수 있고, 추출된 특이음의 데시벨을 측정하여 구간별 Sound score를 도출한다.

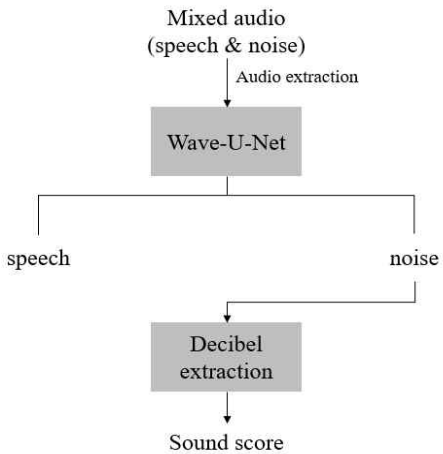

그림 4. Sound score 추출 흐름도
Fig. 4. Extraction flow of sound score

그림 5. Wave-U-Net 학습 결과
Fig. 5. Wave-U-Net training graph

그림 5는 Wave-U-Net 학습 중 혼합 오디오에서 분리한 특이음과 정답 오디오 간의 차이인 L1 Loss가 변하는 그래프이다. 전체 데이터의 20%로 구성된 validation set에 대하여 536k step 이상 학습하는 과정에서 도출되는 Loss의 변화를 확인하였고, 학습이 진행됨에 따라 Loss가 점차 감소하면서 수렴하였다.

4-2 Close-up score 추출 모델

영상에서 얼굴과 손을 검출하고 검출된 영역의 크기 측정을 Yolo v3를 학습하였다. 학습을 위한 데이터셋으로 WIDER FACE[7]와 EgoHand[8]를 사용하였다. WIDER FACE는 얼굴 검출을 위한 데이터셋으로 작은 얼굴부터 큰 얼굴까지 다양한 크기의 얼굴 이미지를 약 32,000장 포함하고 있다. EgoHand 데이터셋은 구글 글래스로 다양한 각도에서 촬영된 4,800장의 손 이미지로 구성된 데이터셋이다.

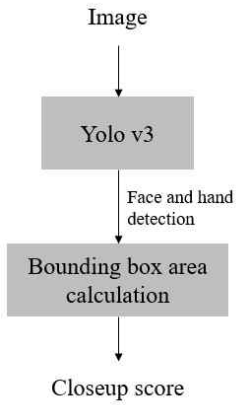

그림 6. Close-up score 추출 흐름도
Fig. 6. Extraction flow of close-up score

WIDER FACE와 EgoHand를 합한 후 이미지의 annotation을 작성하여 YOLOv3 학습을 위한 데이터셋을 만들었다. Yolo v3 학습 후 그림 6과 같이 영상에서 얼굴과 손을 인식하고, 인식된 영역의 bounding box 면적을 계산하여 얼굴과 손의 클로즈업 정도를 측정한다. 측정된 면적의 크기는 Close-up score로 사용한다.

얼굴과 손 검출을 위해 사전훈련된 Yolo v3 모델을 얼굴과 손 데이터셋으로 fine tuning을 진행하였다. 그림 7은 epoch 당 3535 step으로 20 epoch 학습하여 도출된 validation dataset에 대한 결과 그래프이다.

최종 mAP(mean average precision) 0.5는 0.853, mAP 0.5:0.95는 0.563이고 Precision은 0.912, Recall은 0.817의 결과로 학습되었다.

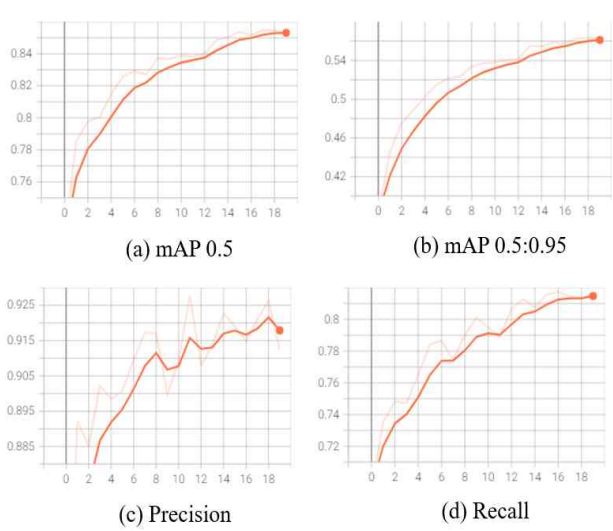

그림 7. Yolo v3 학습 그래프, (a) mAP 0.5, (b) mAP 0.5:0.95, (c) precision, (d) recall
Fig. 7. Yolo v3 training graph, (a) mAP 0.5, (b) mAP 0.5:0.95, (c) precision, (d) recall

4-3 Positivity score 추출 모델

문장의 긍정도를 측정해내기 위해 두 가지 학습 데이터를 사용한다. 첫 번째 데이터셋은 '감성분석을 위한 온라인 상품평 데이터[9]'이다. 이 데이터셋은 화장품 상품평 데이터로 표 3의 형태로 10,000개의 문장이 긍정과 부정으로 분류되어 있다. 본 연구의 텍스트 분석 목적은 입력된 문장이 긍정인지 부정인지 분류하는 것이기 때문에 단문 리뷰로 구성된 데이터셋이 활용에 적합하다. 다음으로 'Naver sentiment movie corpus[10]' 데이터셋을 함께 활용하였다. 이 데이터셋은 표 4 형태의 네이버 영화 리뷰 데이터셋으로 200,000개의 문장이 긍정과 부정으로 분류되어 있다.

위 두 가지 데이터셋을 하나의 데이터셋으로 합하여 LSTM 기반의 binary classification 모델 학습을 하였다.

표 3. 감성분석을 위한 온라인 상품평 샘플데이터[9]
Table 3. Sample Online review sample data for sentiment analysis[9] (The contents of table were written in korean because dataset was derived in korean)

Review	Lable
절대 저런 색상 아니에요 ㅜ 정말 낚임	Negative
생각보다 색상 이쁘지 않아요	Negative
색상 정말 예쁩니다 말 그대로 청순청순	Positive
13호.. 흰끼 더 부각.. 색상 이뻐요	Positive

표 4. Naver sentiment movie corpus 샘플데이터[10]
Table 4. Sample data of Naver sentiment movie corpus[10] (The contents of table were written in korean because dataset was derived in korean)

Review	Lable
이게 영화야?	Negative
정말 재밌게 잘봤습니다!	Positive
롭 코헨의 몰락의 OO점	Negative
감동적이다....	Positive

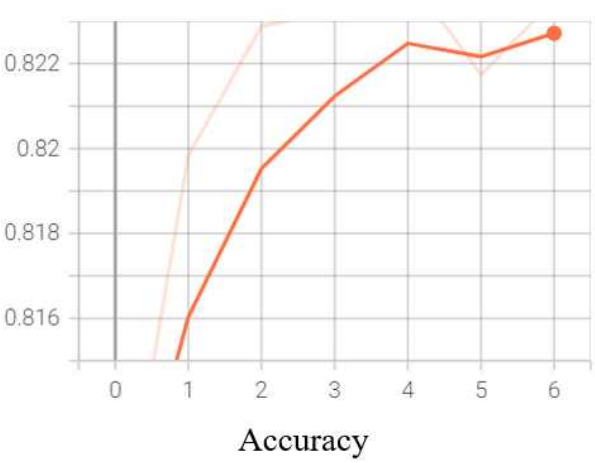

그림 8. Binary classification 모델 학습 그래프
Fig. 8. Binary classification model training graph

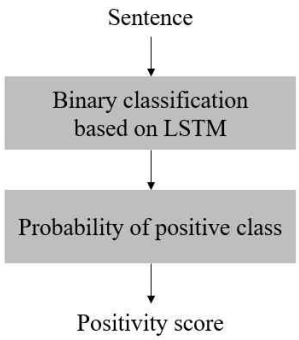

그림 9. 텍스트 임팩트 점수 추출 흐름도
Fig. 9. Text impact score extraction

그림 8은 binary classification 모델의 학습 그래프로 전체 데이터셋의 20%로 구성된 validation dataset에 대한 accuracy 변화를 나타낸다. epoch 당 1915 step으로 학습을 진행하였고 early stopping 방식으로 loss가 높아지면 학습을 종료하여 best accuracy 0.83을 도출하였다.

모델 생성 후 그림 9와 같이 음성을 STT를 통해 텍스트로 변환하고, 해당 문장이 긍정인지 부정인지 분류한다. 그리고 '긍정'일 확률값을 문장(구간)의 Positivity score로 사용한다.

Ⅴ. 평 가

멀티모달 정보 분석을 위해, 4-1, 4-2, 4-3에서 생성한 모델을 가지고 실제 라이브커머스에서 방송된 영상을 대상으로 요약을 진행하였다. 대상 영상은 라이브커머스 플랫폼인 '라이브라떼'에서 2020년 8월 27일에 진행된 방송이고, 약 72분 30초 길이의 원본 영상을 최소 1분의 길이로 요약하였다.

5-1 Sound score 측정

Wave-U-Net으로 음원 분리 후 1초 단위로 특이음의 데시벨 크기를 계산하였다. 그림 10은 3초 단위로 추출된 6개의 오디오를 spectrogram으로 변환한 모습이다. x축은 시간, 과 y축은 주파수를 나타내고 밝기는 데시벨을 의미한다.

a), b), c)와 같이 높은 스코어로 측정된 구간은 영상에서 특이음이 발생한 구간이다. 주로 높은 주파수에서 강한 데시벨을 보인다. 이와는 반대로 d), e), f)와 같이 낮은 스코어로 측정된 구간은 주로 일반적인 톤의 발화 구간이고 높은 스코어 구간 대비 상대적으로 낮은 주파수에서 강한 데시벨을 보인다.

입력 오디오에서 구간별 Sound score를 계산 후 표 5와 같이 약 1분 요약하기 위한 대상 구간을 추출하였다. 대상 구간은 주로 이벤트음, 높은 톤의 음성, 웃음소리를 포함하고 있다.

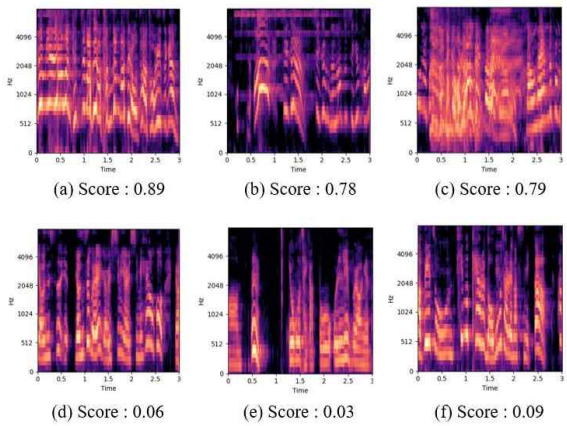

그림 10. 오디오 스펙트로그램, (a) 높은 톤의 음성 (스코어 : 0.89), (b) 이벤트 효과음 (스코어 : 0.78), (c) 웃음소리 (스코어 : 0.79), (d) 일반 음성 1 (스코어 : 0.06), (e) 일반 음성 2 (스코어 : 0.03), (f) 일반 음성 3 (스코어 : 0.09)

Fig. 10. Audio spectrogram, (a) High pitch voice (score : 0.89), (b) Event sound (score : 0.78), (c) Laugh sound (score : 0.79), (d) General voice 1 (score : 0.06), (e) General voice 2 (score : 0.03), (f) General voice 3 (score : 0.09)

표 5. 오디오 스코어 기반 주요 구간 추출

Table 5. Extraction of important video section based on the audio score

Start Time(s)	End Time(s)	Audio Score
14	19	0.94
81	84	0.85
959	963	0.97
1928	1932	0.97
3755	3758	0.85
3761	3765	0.87
3811	3814	0.82
3835	3840	0.88
3849	3853	1.0
3859	3862	0.85
3934	3937	0.83
3979	3982	0.84
4010	4014	0.83
4078	4081	0.93
4202	4205	0.83
4211	4214	0.84
4279	4285	0.89

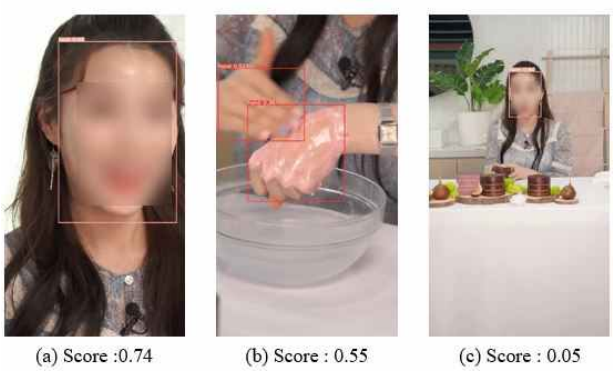

그림 11. 얼굴과 손 인식 결과

Fig. 11. Detection of Face and hand

표 6. 이미지 스코어 기반 주요 구간 추출

Table 6. Extraction of important video section based on the image score

Start Time(s)	End Time(s)	Image Score	ref.
828	838	0.59	Hand close-up
2542	2562	0.95	Face close-up
2791	2810	1.0	Face close-up
3405	3408	0.57	Face close-up
3409	3413	0.49	Face close-up
3418	3426	0.49	Face close-up

5-2 Close-up score 측정

구간별 Close-up score 측정을 위해 영상을 shot 단위로 분할하였다. 프레임 간 픽셀 변화량이 임계치 이상으로 변하면 shot을 구분한다. 주로 카메라 구도의 변화가 있을 때 shot이 분할된다. 다음으로 각 shot의 중앙 프레임을 키프레임으로 선택한 후 얼굴과 손을 검출하는데 사용한다. 키프레임에서 인식된 얼굴과 손의 영역 크기가 해당 구간의 Close-up score가 된다.

그림 11에서 c) 프레임 대비 a), b) 프레임의 Close-up score가 높은 것처럼 얼굴과 손이 인식된 bounding box의 크기를 계산하여 Close-up score를 산출한다. 표6은 Close-up score를 기준으로 대상 구간을 추출한 결과이다. 대상 구간은 주로 제품의 발색을 손등에 바르며 보여주거나, 얼굴에 시연하는 등 제품을 상세히 보여주고 설명하는 구간이 선택되었다.

5-3 Positivity score 측정

Google STT를 통해 영상에서 추출된 오디오를 text로 변환하고, 텍스트에 대한 형태소 분석을 통해 '종결어미'를 기준으로 문장을 분할하여 Positivity score를 측정하였다.

표 7. 텍스트 스코어 기반 주요 구간 추출
Table 7. Extract important video section based text score(The contents of table were written in korean because STT result was derived in korean)

Start Time(s)	End Time(s)	Text Score	ref.
932	942	0.98	그럼요 눈을뜨고 서행 동네는 눈이 또 해 줄 수 있는데 눈을 감고 사진 보시면 누난 뽀 해 주니까 저희 지금 함께 저랑 같이 제품 보면서 프로메디 1열 이게요 I'm from 받아 가실 수 있는 기다리고 있습니다.
1481	1488	0.98	한번 올려 보도록 하고 있거든 너무나 좋겠다
1803	1812	0.98	아 그리고 제가 요거 보고 싶었던 게 에센스 패드는 뭔가 각질이 일어날 때 가든지 일어날 때 굉장히 사용하시면 좋지만 저처럼 오늘 아침 일곱시 에 제가 방정식 나왔었거든요
2595	2602	0.99	하셨는데 지금 이제 조금 있으면 더 예뻐질 건데 생을 예쁘다 감사합니다
2801	2812	0.98	여러분 실제로 저희 지금 배송비도 안 받고 선물로 보내 드리고 있으니까 집에서 잘 사용해 주시면서 이렇게 예쁜 피부 메이크업 또 완성해보세요
3198	3206	0.99	요렇게 손으로 돼 있어 가지고는 굉장히 좋아 곱하기가 좋은 컬러들이 그중에서도 저는 공장에 예쁘더라고요
3219	3227	0.99	너무 예뻐서 다시 한번 들을까 갈색 보여주면서 글리터가 너무 예쁘다고 말씀해주셨는데요
4172	4191	0.98	네 집에 가서 고기 사 드시고 올리고 올리고 그리고 우리 여보 울림 아이고 잘한다 잘한다 영화 보셨어요

표7과 같이 최종적으로 '예쁘다' '좋다'와 같은 긍정적인 의미를 담고 있는 단어 또는 문장이 주로 선정되었다.

5-4 최종 요약 구간 선정

주요 구간 선정을 위해 5-1, 5-2, 5-3에서 도출된 스코어를 합산 후 우선순위에 따라 구간을 선정하였다. 구간 선택 시 영상의 자연스러운 연결을 위해 앞뒤로 1초의 여백을 주어 추출하였고, 최종적으로 표 8과 같이 주요 구간이 도출되었다.

표 8. 멀티모달 점수 합산을 통한 최종 요약 구간 추출
Table 8. Extraction of final important video section using summing each multi-modal score

Start Time(s)	End Time(s)	Total score
2541	2549	2.190668
2551	2554	1.960668
2558	2563	2.010668
2791	2800	2.032435
2802	2811	2.059396
3763	3766	1.904056
3851	3854	1.889339
4098	4123	1.91494

또한, 멀티모달 점수의 가중치를 달리하는 방법으로 요약 구간을 추출하였다.

표 9는 Close-up score의 가중치가 2배일 때 추출된 구간으로 그림 12과 같은 클로즈업 구간에 대한 선택이 높아졌다. 표 10은 Sound score 가중치를 2배로 높였을 때 선정된 대상 구간으로 그림 13과 같이 주로 박수소리, 웃음소리 등 일반 발화의 특징을 벗어난 음이 나오는 구간이 다수 선택되었다. 즉, 멀티모달 스코어의 가중치를 조절하는 방식으로 클로즈업을 중점적으로 보여주거나, 오디오 임팩트에 중점을 두는 등 목적에 따라 요약본을 생성할 수 있었다. 특히 Sound score의 가중치를 높인 경우 방송 중에 진행되는 이벤트 또는 시청자의 몰입도를 높이기 위해 주의를 집중시키는 쇼호스트의 발성 등 라이브커머스 특유의 임팩트 강한 방송 요소들이 효과적으로 요약본에 반영되는 것을 알 수 있었다.

표 9. 멀티모달 점수별 가중치 조절 후 최종 요약 구간 추출 (Positivity score 가중치 : 1.0, Close-up score 가중치 : 2.0, Sound score 가중치 : 1.0)
Table 9. Extraction of final important video section after controlling score weight per multi-modal score (Positivity score weight : 1.0, Close-up score weight : 2.0, Sound score weight : 1.0)

Start Time(s)	End Time(s)	Total score
2541	2554	3.000668
2558	2563	2.960668
2790	2800	3.032435
2802	2843	2.869396

그림 12. 표 9 요약 결과 프레임 (왼쪽 프레임 : Start time(2541), End time(2554), 오른쪽 프레임 : Start time(2790), End time(2800))
Fig. 12. Frame of Table 9 result(Left frame : Start time(2541), End time(2554), Right frame : Start time(2790), End time(2800))

표 10. 멀티모달 점수별 가중치 조절 후 최종 요약 구간 추출 (Positivity score 가중치 : 1.0, Close-up score 가중치 : 1.0, Sound score 가중치 : 2.0)

Table 10. Extraction of final important video section after controlling score weight per multi-modal score (Positivity score weight : 1.0, Close-up score weight : 1.0, Sound score weight : 2.0)

Start Time(s)	End Time(s)	Total score
15	20	2.705755
1464	1467	2.556795
2543	2548	2.610668
3755	3761	2.644056
3762	3766	2.774056
3836	3841	2.759339
3850	3854	2.889339
3860	3863	2.569339
3980	3983	2.612107
4056	4059	2.524891
4079	4082	2.657875
4099	4103	2.51494
4125	4131	2.577134
4213	4216	2.59439
4280	4286	2.537207

그림 13. 표 10 요약 결과 프레임 (왼쪽 프레임 : Start time(15), End time(20), 오른쪽 프레임 : Start time(4280), End time(4286))

Fig. 13. Frame of Table 10 result(Left frame : Start time(15), End time(20), Right frame : Start time(4280), End time(4286))

5-5 영상 요약 결과 및 이전 연구 비교

본 연구의 결과와 Zhou et al. 방식으로 요약한 결과를 비교하였다. 그림 14는 Zhou et al. 방식을 사용한 결과로 선택 프레임 간의 유사성을 낮추는 방식이기 때문에 구간의 다양성은 보이나 Vision 정보만 사용하는 한계로 상호작용과 현장감 같은 라이브커머스 영상의 특징이 부족했다. 반면, 본 연구의 결과는 그림 15와 같이 오디오 정보 분석으로 시청자와의 상호작용 및 현장감이 높은 구간을 검출하고 이미지와 텍스트를 함께 분석하여 제품을 상세하게 확인할 수 있는 구간을 찾아냄으로써 라이브커머스 방송의 중요한 요소가 다양하게 포함되었다.

그림 14. Zhou et al. 방식의 영상 요약 샘플 프레임

Fig. 14. Sample frame of video summarization using method of Zhou et al.

그림 15. 본 논문 방식으로 요약한 영상의 샘플 프레임

Fig. 15. Sample frame of video summarization using method of this paper

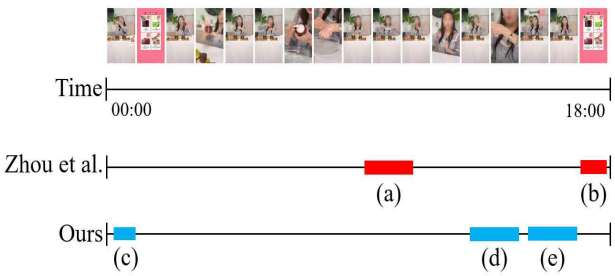

그림 16. 첫 번째 상품 판매 영상에서의 구간 추출 비교 ((a) 일반 멘트 (12:43 ~ 12:45), (b) 일반 멘트 (17:40 ~ 17:41), (c) 인사말, 박수 (00:16 ~ 00:18), (d) 제품 상세 시연 (13:52 ~ 13:56), (e) 이벤트 (16:01 ~ 16:03))

Fig. 16. Comparison of section extraction from the first product sales video ((a) general comment (12:43 ~ 12:45), (b) general comment (17:40 ~ 17:41), (c) greeting, clap (00:16 ~ 00:18), (d) detail trial performance of product (13:52 ~ 13:56), (e) event (16:01 ~ 16:03))

그림 16은 전체 영상 중 첫 번째 상품을 판매하는 구간을 대상으로 본 연구의 결과와 Zhou et al. 방식으로 요약한 결과를 비교한 모습이다.

Zhou et al. 방식으로 선택된 구간은 일반적인 멘트로 진행되는 구간이 뽑힌 것과 다르게 본 논문의 결과에서는 높은 음성 피치와 데시벨로 인사하며 박수치는 구간과 제품을 손등에 시연한 후 상세하게 보여주는 구간, 이벤트가 시작되는 구간이 추출된다. 영상을 풍부하게 해석하기 위해 멀티모달 정보를 활용함으로써 임팩트 강한 구간에 집중하여 요약본을 생성하는 것을 확인하였다.

Ⅴ. 결 론

본 연구에서는 라이브커머스 방송의 요약본 VOD 서비스 제공을 위한 자동 영상 요약 기법을 제안하였다. 제안한 방법은 오디오, 이미지, 텍스트로 구성된 세 가지 멀티모달 정보를 딥러닝 기반으로 분석하여 영상 중 임팩트가 강한 구간을 선택하는 방법으로 영상을 요약한다. 실험을 통해 Wave-U-Net을 활용한 음원 분리로 특이음을 분리하여 임팩트 강한 구간을 추출하였고, 키프레임에서 얼굴과 손을 인식하여 클로즈업되는 구간을 검출하였다. 또한 출연자의 음성을 STT로 변환한 후 텍스트의 긍정도를 측정하여 출연자가 긍정적인 멘트로 발화하는 구간을 추출하였다. 한편 문장 긍정도 분석 과정에서 STT로 변환된 텍스트가 출연자의 스크립트를 모두 반영하지 못했음에도 불구하고 문장의 맥락을 기반으로 긍정도를 측정해냄으로써 Positivity score를 계산하였다. 최종적으로 라이브커머스 방송의 생동감과 주요 이벤트, 제품 설명 등 주요 구간이 담겨 있는 요약 결과를 확인하였다.

본 기법으로 실시간 라이브커머스 방송 종료 직후 영상을 자동 요약함으로써 요약본 제공에 대한 적시성 확보와 영상 편집에 필요한 시간과 비용에 대한 단축 효과를 기대할 수 있다. 또한 각 멀티모달 정보 분석 후 도출된 스코어의 가중치를 조절함으로써 사용자의 의도가 반영된 요약본 생성이 가능하였다.

향후 더욱 효과적인 주요 구간 검출을 위해 스크립트에 대한 STT 변환 정확도를 개선하고자 한다. 또한, 요약 영상의 목적에 따라 멀티모달 스코어 가중치를 자동화된 형태로 설정할 수 있도록 연구를 확장하여, 이를 토대로 영화 또는 드라마, 스포츠 등 다양한 영역에 적용함으로써 딥러닝을 활용한 자동 영상 요약 분야에 공헌하고자 한다.

참고문헌

[1] Kaiyang Zhou, Yu Qiao, and Tao Xiang, “Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward”, in *Association for the Advancement of Artificial Intelligence*, New Orleans, pp 7582–7589, 2018.

[2] M. Rochan, L. Ye, and Y. Wang, “Video summarization using fully convolutional sequence networks”, in *European Conference on Computer Vision*, Munich, pp. 358–374, 2018. doi.org/10.1007/978-3-030-01258-8_22

[3] A. R. Lee, “Investigating the Factors Influencing the Use of Live Commerce in the Un-tact Era: Focusing on Multidimensional Interactivity, Presence, and Review Credibility”, *Knowledge Management Review*, Vol. 22, No. 1, pp. 269-286, 2021. doi.org/10.15813/kmr.2021.22.1.013

[4] Daniel Stoller, Sebastian Ewert, and Simon Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation”, in *International Society for Music Information Retrieval Conference*, Paris, pp. 334–340, 2018.

[5 Joseph Redmon and Ali Farhadi, “YOLOv3: An incremental improvement”, arXiv preprint arXiv:1804.02767, 2018.

[6] David Snyder, Guoguo Chen, and Daniel Povey, “Musan:A music, speech, and noise corpus”, arXiv preprint arXiv:1510.08484, 2015.

[7] S. Yang, P. Luo, C.-C. Loy, and X. Tang, “Wider face: A face detection benchmark”, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu, “Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions”, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1949–1957, 2015.

[9] Shin, K-s. Data Repository for Business Research [http://www.drbr.or.kr]. The Korean Academic Society of Business Administration. 2019
[10] Naver sentiment movie corpus v1.0 [Internet]. Available : https://github.com/e9t/nsmc

황중수(Jung-Su Hwang)

2014년 : 성균관대학교 컴퓨터공학과 (공학사)
2017년~현 재: CJ올리브네트웍스 AI Core 연구소
※관심분야 : 인공지능, 영상요약, 미디어콘텐츠 분석 등

조영선(Young-Sun Cho)

2015년 : 한양대학교 대학원 (공학석사)
2018년~현 재: CJ올리브네트웍스 AI Core 연구소
※관심분야 : 인공지능, 영상요약, 이미지 처리 등

박상훈(Sang-Hoon Park)

2016년 : 가톨릭대학교 대학원 (공학석사)
2018년 : 가톨릭대학교 대학원 (공학박사-Brain Computer Interface)
2018년~2020년: 롯데정보통신 정보기술연구소
2021년~현 재: CJ올리브네트웍스 AI Core 연구소
※관심분야 : 인공지능, 음성인식, 인간-컴퓨터 상호작용 등

이현기(Hyun-Ki Lee)

2014년~현 재 : CJ올리브네트웍스
※관심분야 : 인공지능, 얼굴합성, 영상요약 등

손종수(Jong-Soo Sohn)

2003년 : 고려대학교 대학원 (이학석사)
2013년 : 고려대학교 대학원 (이학박사-지능컴퓨터시스템)
2013년~2020년: 삼성전자 VD사업부 컨텐츠 추천엔진 PL
2020년~현 재: CJ올리브네트웍스 AI Core 연구소장
※관심분야 : AI, Deep Learning, 컨텐츠 추천 등